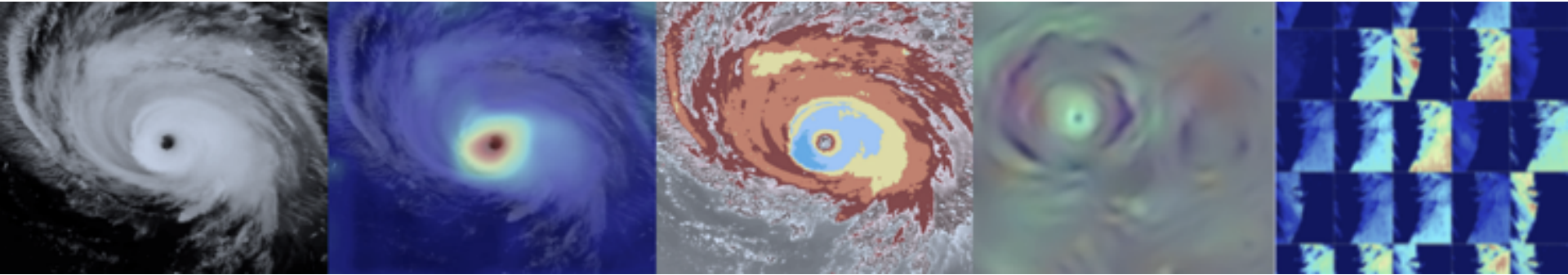


Advancing AI for Earth Science: A Data Systems Perspective



Manil Maskey, Ph.D.

NASA

Earth Science Data Systems/NASA Headquarters

Interagency Implementation and Advanced Concepts Team (IMPACT)/NASA Marshall Space Flight Center

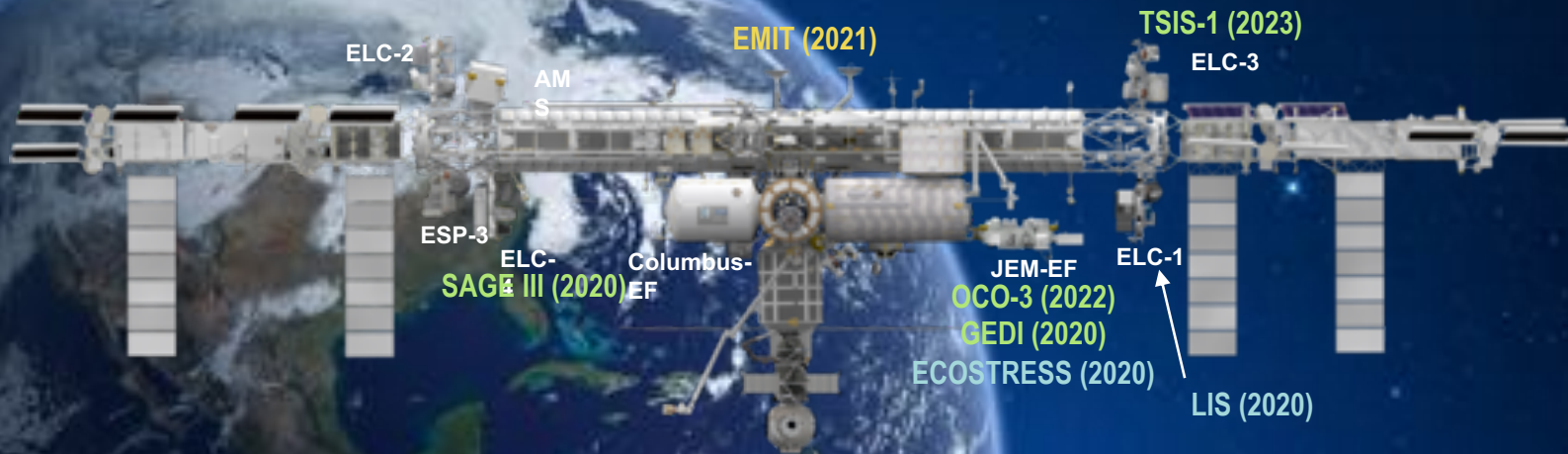
NASA EARTH FLEET

OPERATING & FUTURE THROUGH 2023



INTERNATIONAL SPACE STATION

EARTH SCIENCE OPERATING MISSIONS



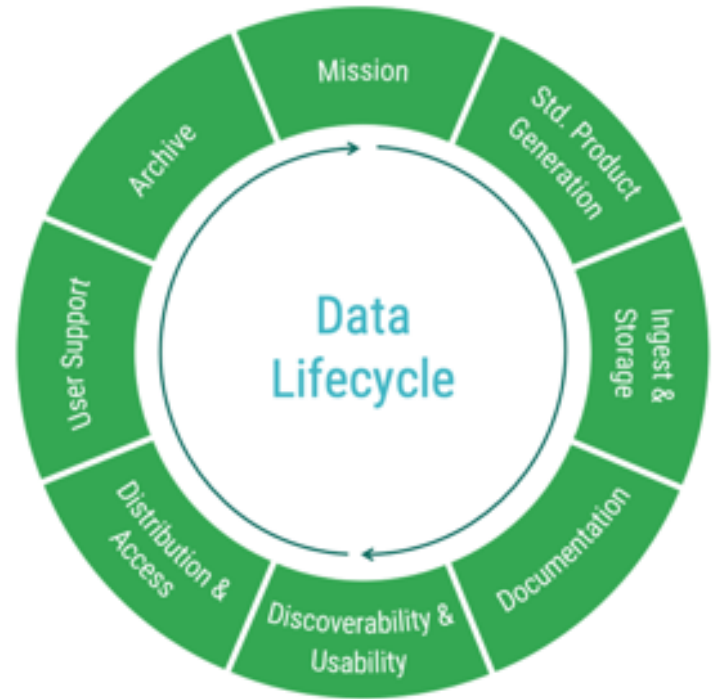
EXPRESS Logistics Carriers: ELC-1, ELC-2, ELC-3
External Stowage Platforms: ESP-3
Alpha Magnetic Spectrometer: AMS
Columbus External Payload Facility: Columbus-EF
Kibo External Payload Facility: JEM-EF

(PRE) FORMULATION ●
IMPLEMENTATION ●
PRIMARY OPS ●
EXTENDED OPS ●

NASA's Earth Science Data Systems Program

Single largest repository of Earth Science Data, integrating multivariate/heterogeneous data from diverse observational platforms

Manages NASA's Earth science data through the entire data life cycle



Why is this important?

“The fraction of science papers that rely on archive data is increasing and, in many cases, exceeds the fractions of papers based on new mission data.” - NASA Advisory Council Ad-Hoc Big Data Task Force



Enabled by 25+ years of

Open Data
Open Source
Open Services



ESDS by the numbers –FY19

Unique data products

34,500

Data products distributed

1.9 billion

Archived files

462 million

Current archive

33.6 PB

Distinct users

3.5 million

American Customer Satisfaction Index

78





Challenges

- Prepare for planned high-data-rate missions
- Improve efficiency of NASA's data systems operations
- Increase opportunity for researchers and commercial users to access/process PBs of data quickly without the need for data management
- Transparent/extendable open source processing framework
- Ensure users find right data for their problem
- Minimize user burden to access data
- Enable users to extract new knowledge/information from archives



Enabling technologies

Cloud computing

- Operate components of the data systems in a commercial cloud environment to meet future needs
- Provide new opportunities for users to process data in place and perform analytics at scale

Data-driven technologies

- Maximize information and knowledge discovery capabilities
- Augment data stewardship processes
- Address Earth science research and application needs



“Big Data Close to Compute”

Large Volume Data Storage

Centralized mission observation & model datasets stored in auto graduated AWS object storage (S3, S3-IA, Glacier)

Applications & Services

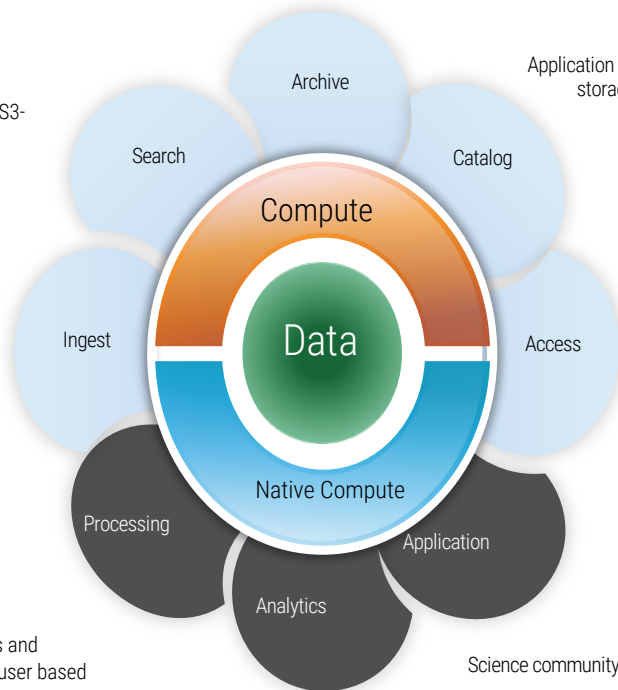
Application and service layer using AWS compute, storage (S3, S3IA, Glacier), and cloud native technologies

Scalable Compute

Provision, Access, and terminate dynamically based on need. Cost by use

Cloud Native Compute

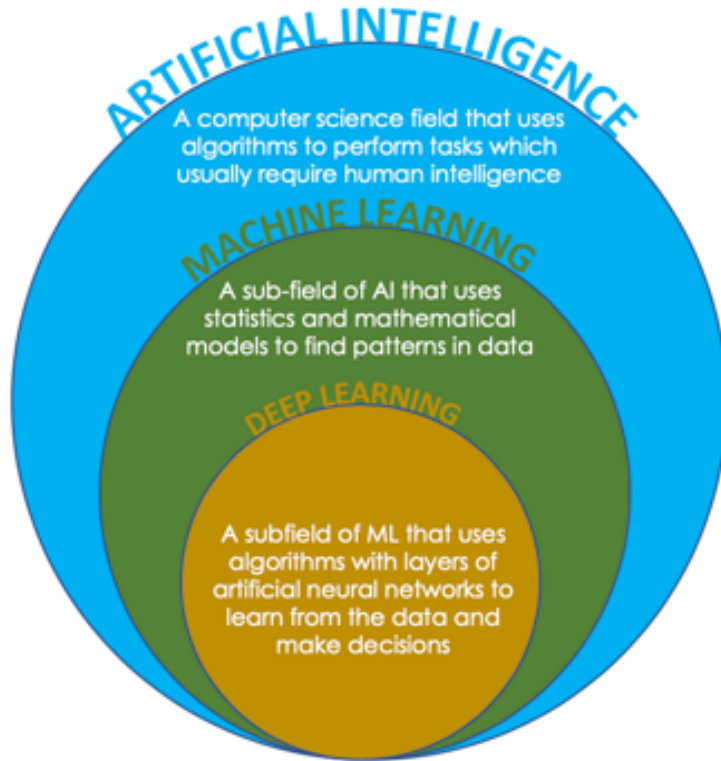
Cloud vendor service software stacks and microservices easing deployment of user based applications



Public Applications & Services

Science community brings algorithms to the data. Support for NASA & non-NASA

Data driven technologies



Rapid adoption of AI/ML due to:

Expanding data volumes

Improving Algorithms

Networks

Cloud computing

Hardware

Maximize information and knowledge
discovery capabilities

Phenomena portal



Why?

Increasing Earth science data archives require non-traditional approaches to data management

Data driven technologies to provide advanced search capabilities

Machine learning-based approach - an enabling data driven technology to provide automated detection of Earth science events from image archives

Catalog of events can provide a novel way to explore large archives of data

Discover and explore Earth science data archives around events using machine learning (ML) techniques



VISIBLE LAYERS

- Chlorophyll a
[On] [Off] [Settings]
- Coastlines / Borders / Roads
[On] [Off] [Settings]
- Global Precipitation from Space
[On] [Off] [Settings]
- Sea and Thermal Anomalies (Sea)
[On] [Off] [Settings]
- SeaWiFS SST
[On] [Off] [Settings]

BASE LAYERS

- Continental Shelfline (Blue Line)
[On] [Off] [Settings]
- Continental Shelfline (Pink Line)
[On] [Off] [Settings]

All Layers | Base Comparison

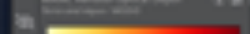


FIRE LAYERS

Fire Perimeter
US Department of Agriculture, National Fire Plan

Capitons / Borders / Roads
US Department of Agriculture, National Fire Plan

Global Advanced Very High Resolution
MODIS/VIIRS V2000



Cloud and Thermal Anomalies (Day
Offsets)

Temperature Anomalies

NO2 Layers

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000

Global Advanced Very High Resolution
MODIS/VIIRS V2000





13 Matching Collections

Sort by: Relevance Only include collections with granules Include non-EODSIS collections

Tip: Add collections to your project to compare and download their data.



NOAA/NASA Thermal Anomalies/Sea & Day L2 Global Sea SW Grid V006

4 Granules • 2000-02-16 ending • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Sea & Day (M2D1442) Version 6 data are generated at 1 kilometer (km) spatial resolution as a Level 2 product. The M2D1442 gridded composite contains the maximum value of individual five pixel values detected during the eight days of acquisition. The Science Dataset (SDS) layers include the file name, pixel quality indicators, Improvement/Changes from Previous Versions, ...

[View this collection](#)



NOAA/NASA Thermal Anomalies/Sea Daily L2 Global Sea SW Grid V006

4 Granules • 2000-02-16 ending • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Sea Daily (M2D1443) Version 6 data are generated every eight days at 1 kilometer (km) spatial resolution as a Level 2 product. M2D1443 contains eight consecutive days of the data consistently packaged into a single file. The Science Dataset (SDS) layers include the file name, pixel quality indicators, maximum the relative power (MaxRPS), and the position of the five pixel within the sea...

[View this collection](#)



NOAA/NASA Thermal Anomalies/Sea Daily L2 Global Sea SW Grid V006

4 Granules • 2000-07-04 ending • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Sea Daily (M2D1443) Version 6 data are generated every eight days at 1 kilometer (km) spatial resolution as a Level 2 product. M2D1443 contains eight consecutive days of the data consistently packaged into a single file. The Science Dataset (SDS) layers include the file name, pixel quality indicators, maximum the relative power (MaxRPS), and the position of the five pixel within the sea...

[View this collection](#)



NOAA/NASA Thermal Anomalies/Sea & Day L2 Global Sea SW Grid V006

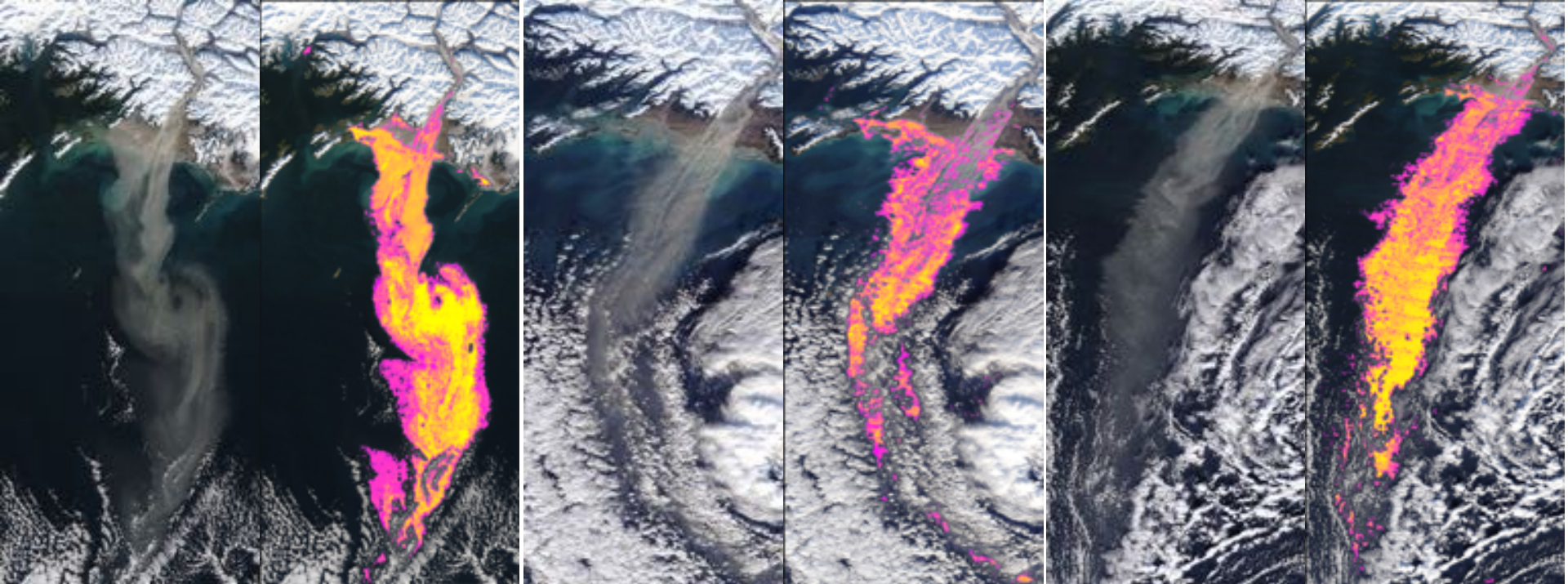
4 Granules • 2000-07-04 ending • The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Thermal Anomalies and Sea & Day (M2D1442) Version 6 data are generated at 1 kilometer (km) spatial resolution as a Level 2 product. The M2D1442 gridded composite contains maximum value of individual five pixel values detected during the eight days of acquisition. The Science Dataset (SDS) layers include the file name, pixel quality indicators, Improvement/Changes from Previous Versions, ...

[View this collection](#)



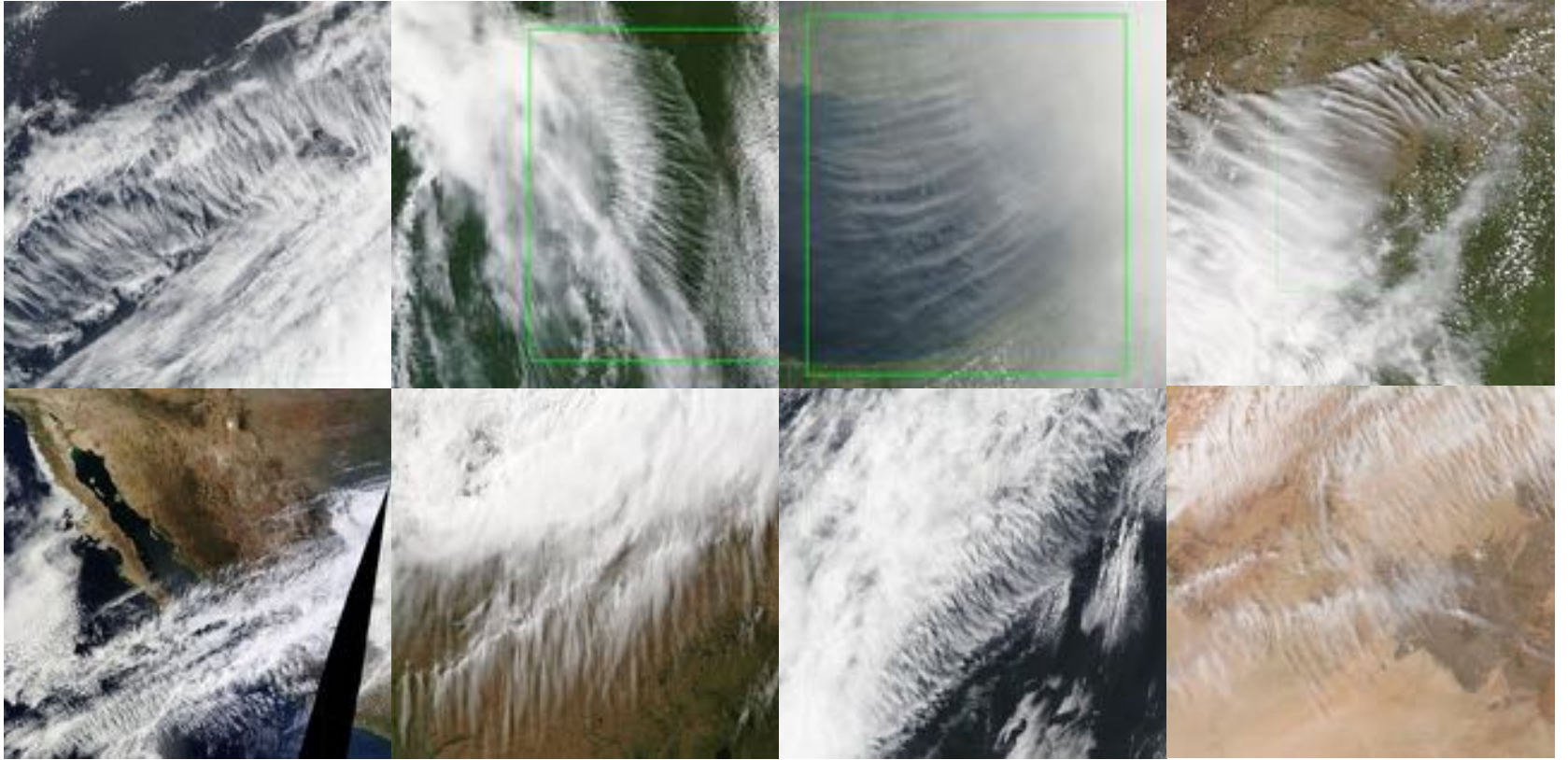
Kincade Fire Demo Video





High latitude dust

Transverse cirrus bands



Phenomena Portal Demo Video

Augment data stewardship processes

Automated keyword assignment



Why?

Assigning science keywords is currently a manual process, which is prone to human error and inconsistencies.

Metadata managed across a network of multiple data centers (i.e. keywords not assigned by a central entity)

Keywords may be assigned by non-subject matter experts (SMEs)

Improve metadata quality

Provide objective and consistent approach to keyword assignment

LIS/OTD 2.5 Degree Low Resolution Annual Climatology Time Series (LRACTS) V2.3.2015

CMR Dataset Title and Description

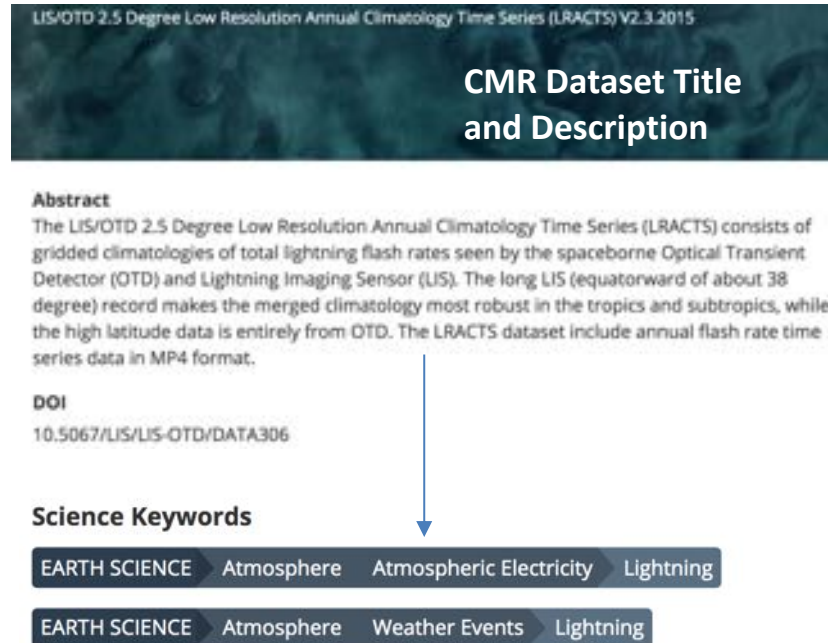
Abstract
The LIS/OTD 2.5 Degree Low Resolution Annual Climatology Time Series (LRACTS) consists of gridded climatologies of total lightning flash rates seen by the spaceborne Optical Transient Detector (OTD) and Lightning Imaging Sensor (LIS). The long LIS (equatorward of about 38 degree) record makes the merged climatology most robust in the tropics and subtropics, while the high latitude data is entirely from OTD. The LRACTS dataset include annual flash rate time series data in MP4 format.

DOI
10.5067/LIS/LIS-OTD/DATA306

Science Keywords

EARTH SCIENCE Atmosphere Atmospheric Electricity Lightning

EARTH SCIENCE Atmosphere Weather Events Lightning



Approach – build word embeddings

Journal Name	Date Published																	
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Atmospheric Science Letters		5	21	34	27	22	42	39	39	44	34	80	69	67	99	61	67	53
Earth and Space Science													1	24	28	25	42	88
Earth's Future												13	26	29	32	31	89	56
Eos, Transactions American Geophysical Union										46	37			1				
Geochemistry, Geophysics, Geosystems	86	202	246	373	355	343	328	374	344	283	292	247	336	354	385	337	64	29
Geoshealth																22	34	14
Geophysical Research Letters	1,322	1,436	1,558	1,696	1,709	1,513	1,509	1,290	1,058	1,204	1,026	1,154	1,266	1,389	1,481	1,390	1,507	1,338
Global Biogeochemical Cycles	137	136	124	128	75	46	94	66	83	44	36	76	77	26	75	74	23	21
Journal of Advances in Modeling Earth Systems							6	3	3	26	25	49	88	56	113	125	139	
Journal of Geophysical Research						36				20								
Journal of Geophysical Research: Atmospheres	175	1,238	786	756	384	563	722	741	969	945	280	76	117	238	311	256	281	378
Journal of Geophysical Research: Biogeosciences				23	29	140	117	117	146	145	130	92	103	133	138	132	45	24
Journal of Geophysical Research: Earth Surface		52	47	31	84	145	130	113	134	132	141	117	92	92	93	58	44	30
Journal of Geophysical Research: Oceans	254	492	325	314	312	341	434	323	302	521	428	338	343	330	305	308	492	269
Journal of Geophysical Research: Planets	117	279	127	126	169	159	195	130	142	171	122	202	89	28	25	103	142	92
Journal of Geophysical Research: Solid Earth	345	603	465	315	372	435	436	376	589	460	328	297	315	362	354	308	326	46
Journal of Geophysical Research: Space Physics	428	543	436	525	475	442	505	533	256	271	494	490	503	543	593	542	466	446
Metereological Applications							7	42	60	26	16	46	61	67	76	1	2	
Paleoclimatology	61	109	98	62	42	41	81	47	64	59	55	45	59	26	54	54		
Paleoclimatology and Paleobotany																	24	26
Quarterly Journal of the Royal Meteorological Society									6	105		203	168	192	128			
Radio Science	108	132	146	114	94	122	91	208	59	136	76	52	63	29	109	91	94	51
Reviews of Geophysics	9	23	12	11	9	34					12	12	14	22	36	22	16	17
Space Weather		16	17	13	48	47	44	46	88	55	43	66	53	62	48	88	125	58
Tectonics	33	45	78	88	58	73	46	65	73	47	28	58	74	83	99	116	43	59
Water Resources Research	302	296	319	317	328	364	314	254	250	404	408	442	412	364	403	337	413	346

88,410
documents

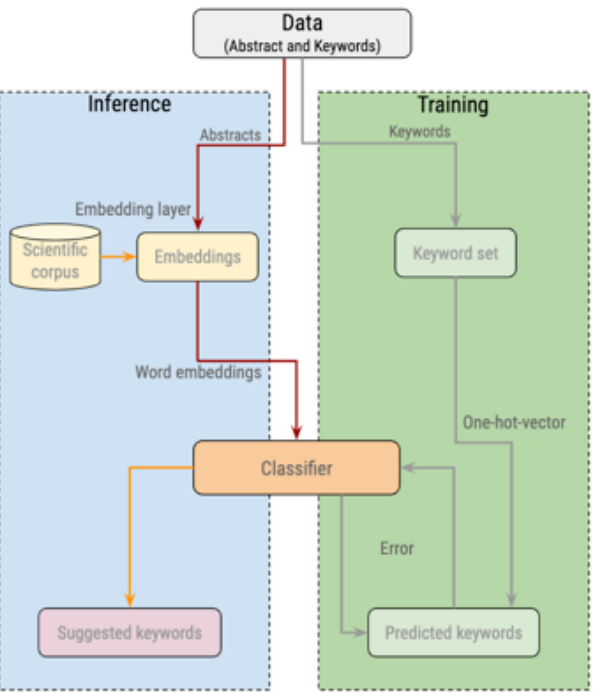
530 million
words

5.5 million
unique words

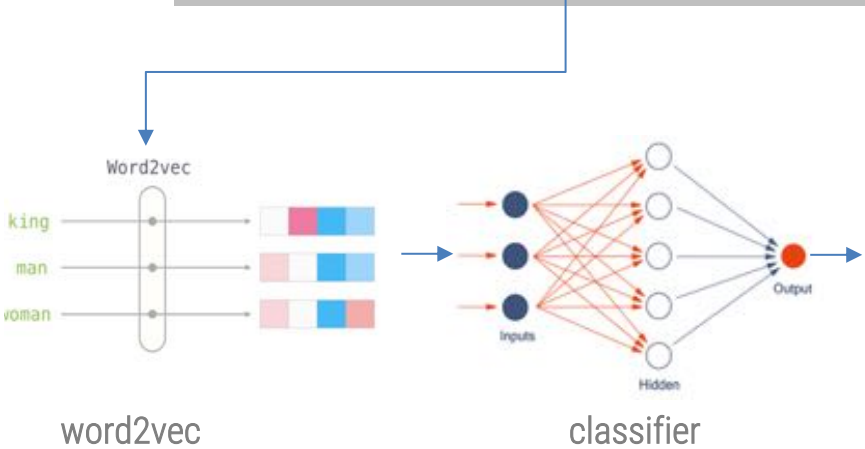


Automated keyword assignment

Version 7.3 is the current version of the data set. Version 3.5 is no longer available and has been superseded by Version 7.3. This data set is currently provided by the OCO (Orbiting Carbon Observatory) Project. In expectation of the OCO-2 launch, the algorithm was developed by the Atmospheric CO2 Observations from Space (ACOS) Task as a preparatory project, using GOSAT TANSO-FTS spectra. After the OCO-2 launch, "ACOS" data are still produced and improved, using approaches applied to the OCO-2 spectra. The "ACOS" data set contains Carbon Dioxide (CO2) column averaged dry air mole fraction for all soundings for which retrieval was attempted. These are the highest-level products made available by the OCO Project, using TANSO-FTS spectral radiances, and algorithm build version 7.3. The GOSAT team at JAXA produces GOSAT TANSO-FTS Level 1B



→ Training flow → Inference flow → Training and Inference flow



Predicted Keyword	Score
carbon dioxide	0.4513424
land use/land cover classification	0.3825603
terrain elevation	0.1924277
barometric altitude	0.18085223
carbon and hydrocarbon compounds	0.07634798

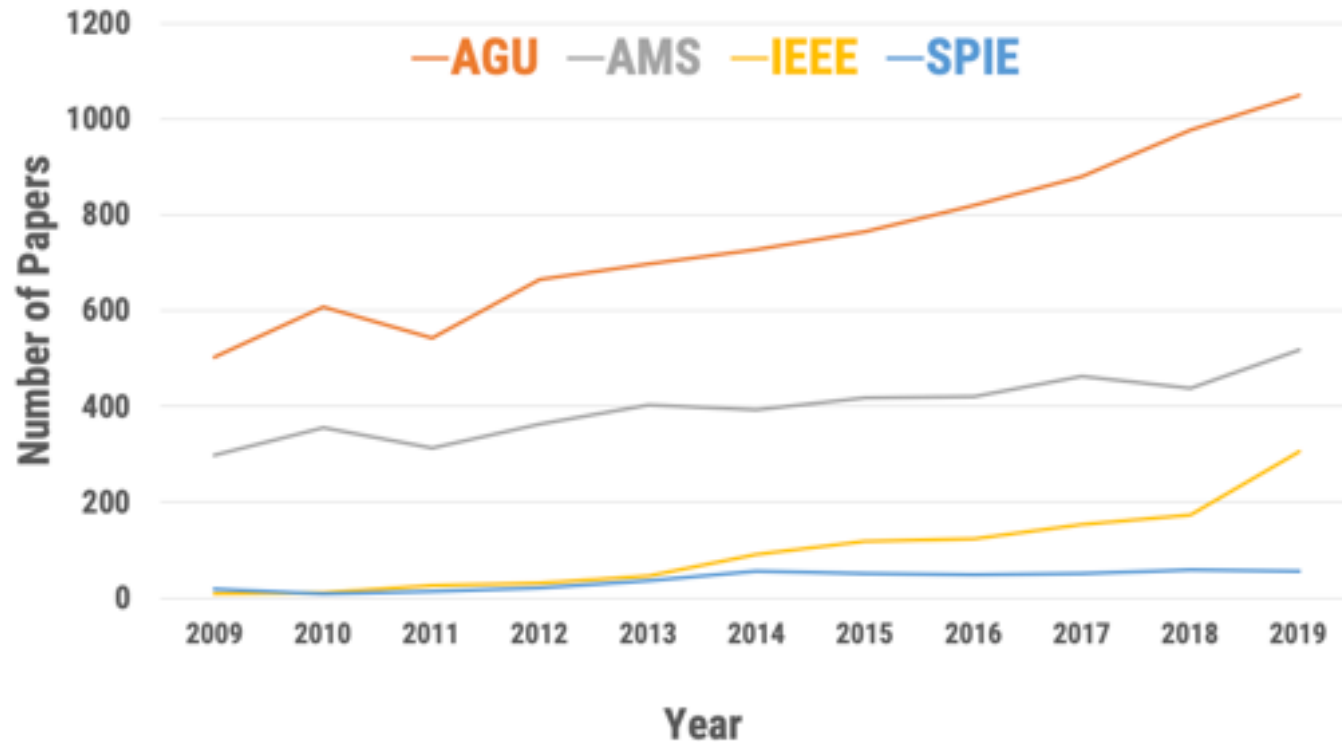


Address Earth science research and
application needs:

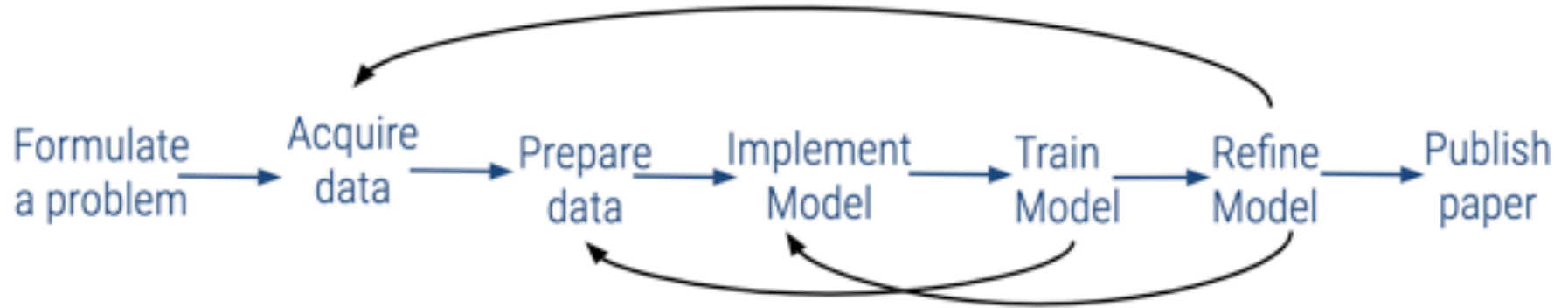
Hurricane intensity estimation system



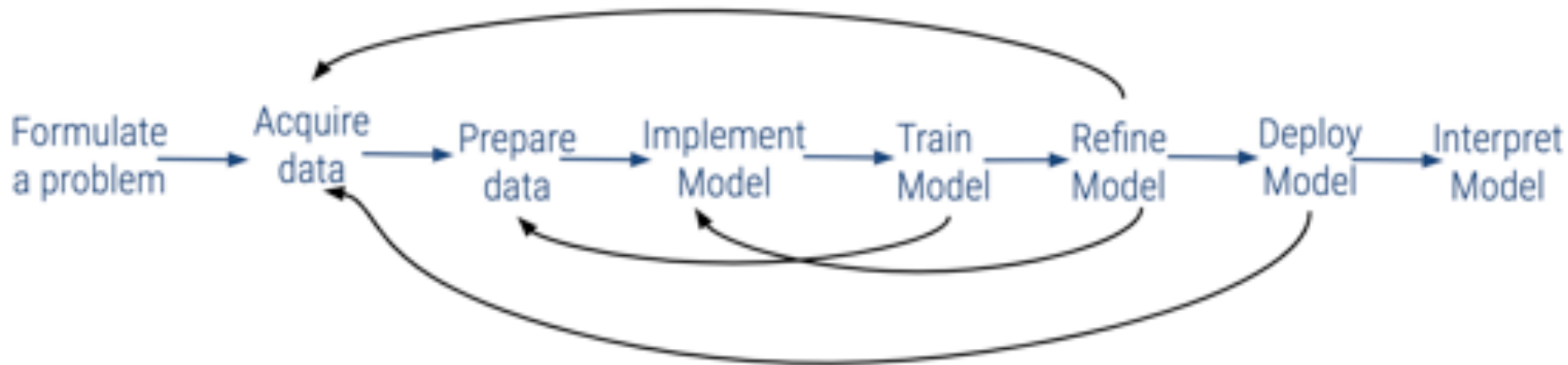
AI/ML in Earth Science



ML in literature



ML lifecycle - iterative



Motivation

15 UTC 10 Oct 17 NHC advisory on Tropical Storm Ophelia:

“Dvorak intensity estimates range from T2.3/33 kt from UW-CIMSS to T3.0/45 kt from TAFB to T4.0/65 kt from SAB. For now, the initial intensity will remain at 45 kt, which is an average of the scatterometer winds and all of the other available intensity estimates.”



Motivation

15 UTC 10 Oct 17 NHC advisory on Tropical Storm Ophelia:

*“Dvorak intensity estimates range from T2.3/**33 kt** from UW-CIMSS to T3.0/**45 kt** from TAFB to T4.0/**65 kt** from SAB. For now, the initial intensity will remain at **45 kt**, which is an average of the scatterometer winds and all of the other available intensity estimates.”*



Motivation

15 UTC 10 Oct 17 NHC advisory on Tropical Storm Ophelia:

“Dvorak intensity estimates range from T2.3/33 kt from UW-CIMSS to T3.0/45 kt from TAFB to T4.0/65 kt from SAB. For now, the initial intensity will remain at 45 kt, which is an average of the scatterometer winds and all of the other available intensity estimates.”

Can we objectively estimate wind speed from satellite images?

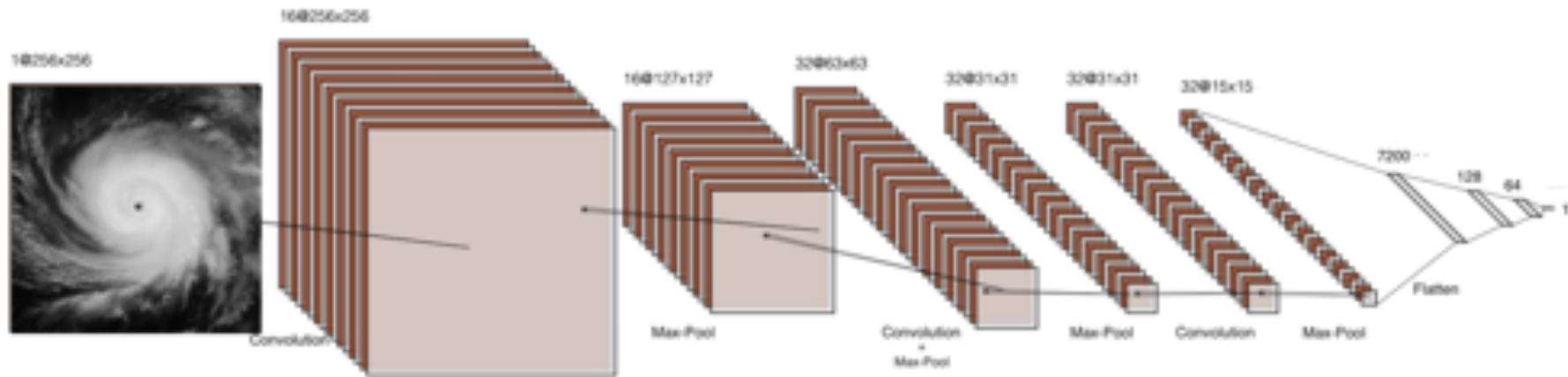
Can we estimate more frequently?



Data

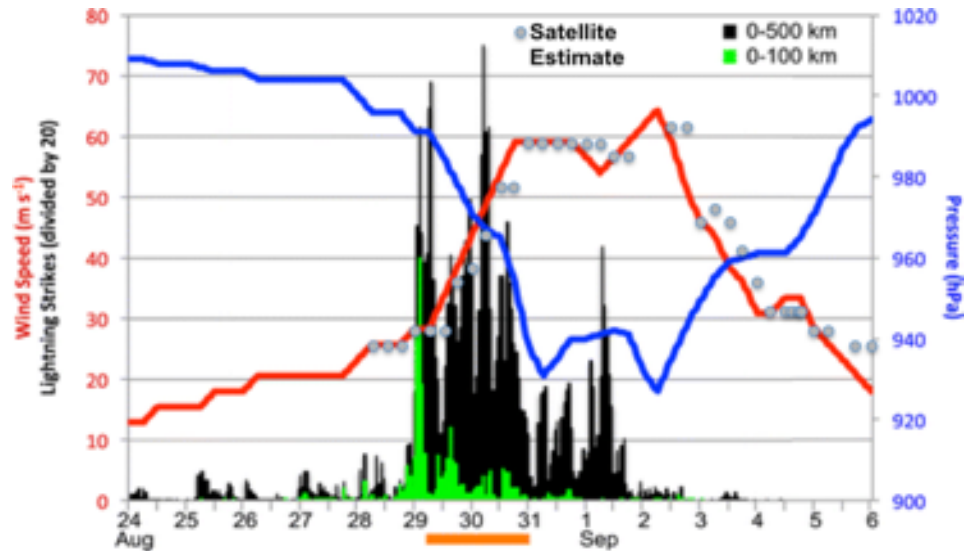


Model development



Test results

Detailed look: Hurricane Earl, 2010



Adapted from Stevenson et al. (2014). Time series of satellite-derived intensity estimates (circles) for Hurricane Earl (2010), added to best track intensities and lightning flash rate time series.

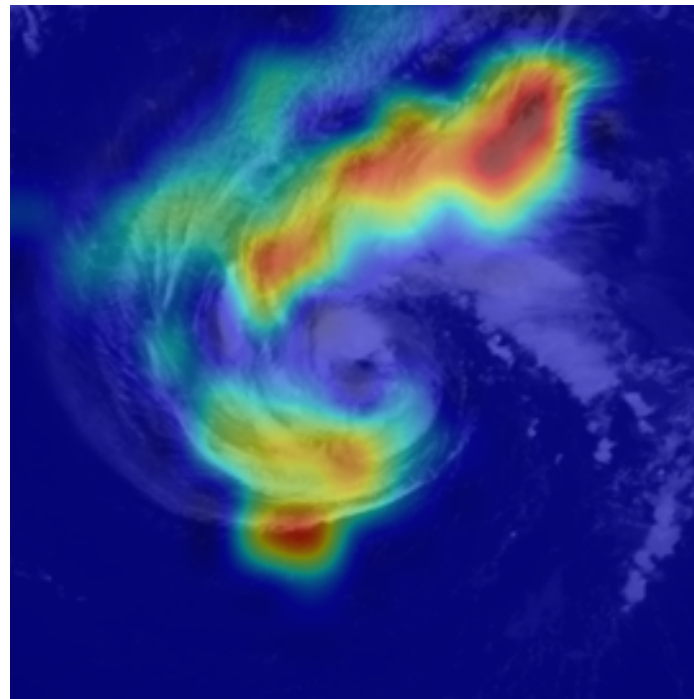
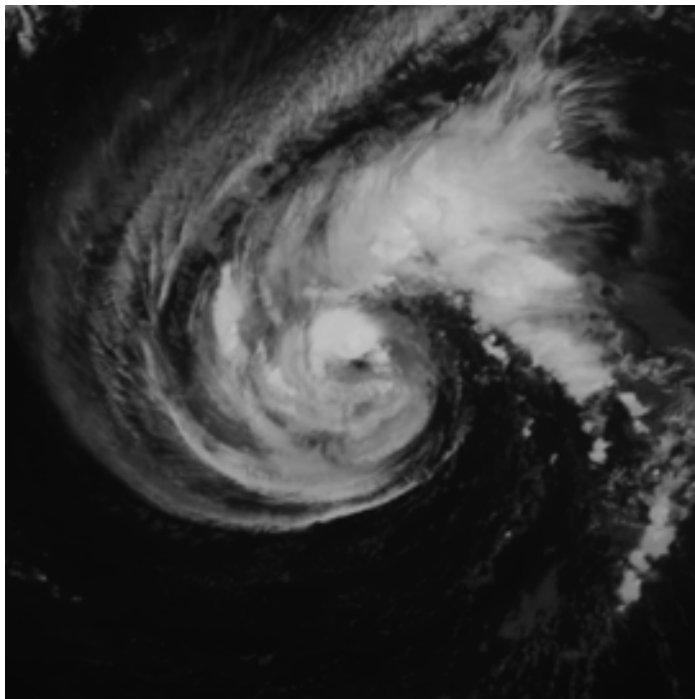
AI as a black box

Interpretability + model inspection



Learning Deep Features for Discriminative Localization

Model evaluation with class activation maps



Tropical Storm

Hurricane Dorian Demo Video

We have a model....now what?

Going extra mile

Interpret prediction data – prediction output maybe just numbers

Questions:

Does the model confidence remain the same over time?

How do you maintain?

How do you complete the loop with new training data?



Deployment to production

Performance requirements

Metrics and baselines with initial models

Monitor over time

Back-testing

Model and software will change

Testing model changes on historical data

Run current production model to baseline performance

Run new models, competing for production

Now-testing

Testing of production model on latest data

Can we get early warning that the model may be faltering?

- Content drift: training data exploited by model are subtly changing with time



Hurricane wind speed estimation portal

Features of a situational awareness tool:

Monitor NHC outlook for “invest” area for trigger

Near real-time tropical cyclone intensity estimation services

Map display

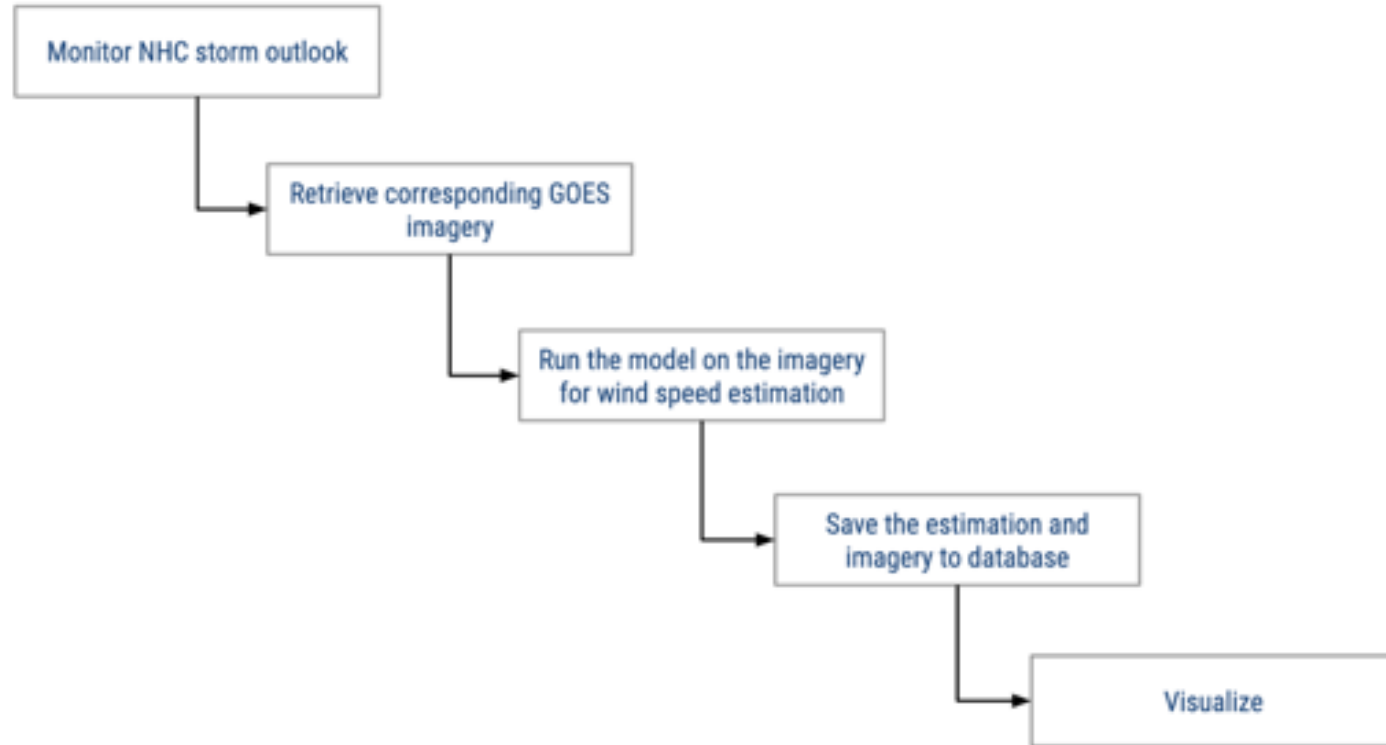
Layers

Comparison with operational forecasts/Evaluate

Service APIs



Workflow



Coordinated effort

ML researchers

- Transform ideas into models
- Training data
- Monitor

Domain experts

- Evaluation
- Performance baselines
- Science use case

End-user stakeholders

- Production requirements

ML engineers

- Design
- Quick prototype
- Deployment
- Scale
- Log



Hurricane Intensity Estimation Portal Demo Video

Challenges and lessons learned

Consistent large-scale training data

AI black box

Training data/Input data becomes part of the code

Versioning training data, model, algorithm becomes difficult

DevOps, CI/CD

Complexity with evolving platforms and infrastructure

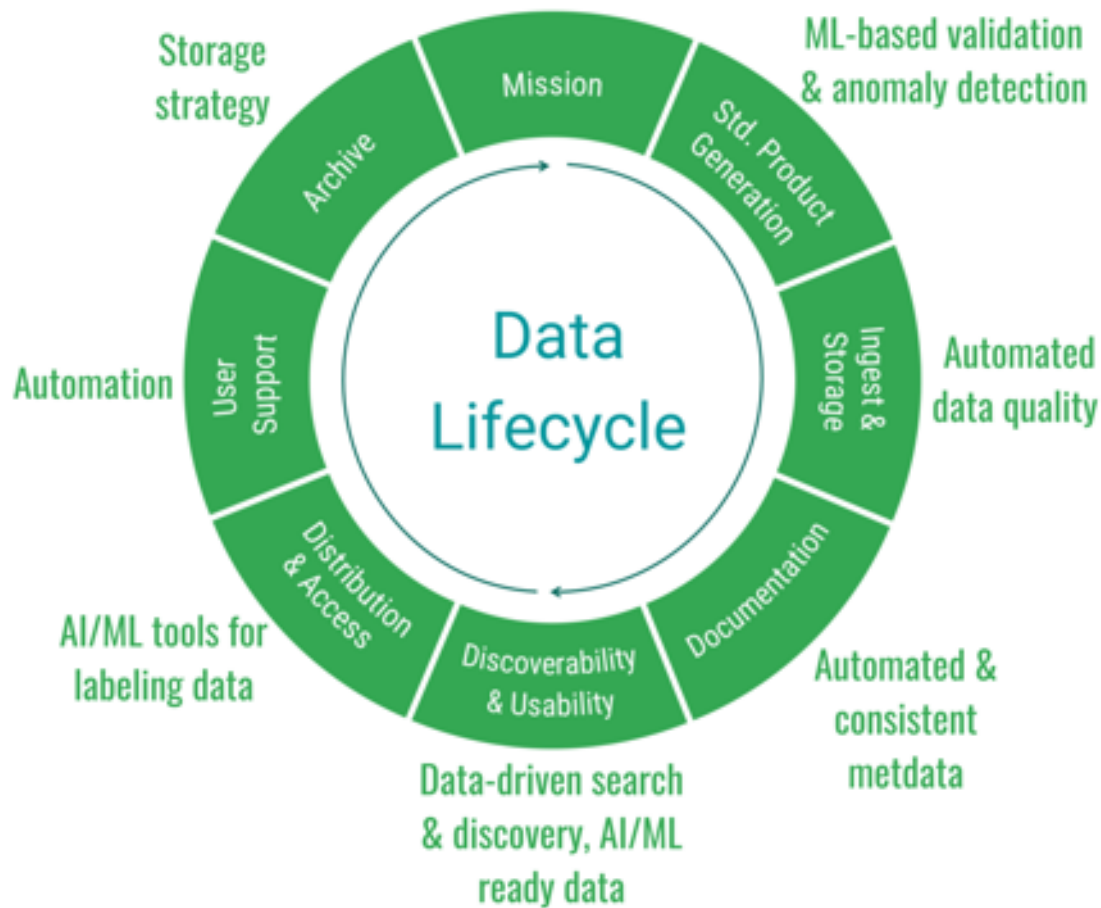


What's next?



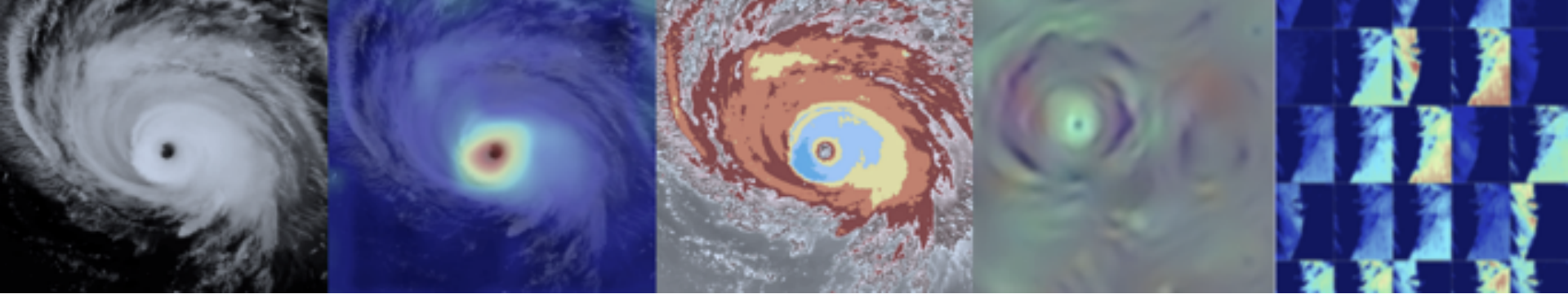
Biggest bottleneck in adoption of ML in Earth
Science is training data





AI for full data lifecycle





Thank you.

Manil Maskey
manil.maskey@nasa.gov

